Environ Ecol Stat (2008) 15:241–258
DOI 10.1007/s10651-007-0049-5

A Bayesian approach to multi-source forest area estimation

Andrew O. Finley · Sudipto Banerjee · Ronald E. McRoberts

Received: 12 October 2005 / Revised: 14 June 2006 / Published online: 25 October 2007

Abstract In efforts such as land use change monitoring, carbon budgeting, and forecasting ecological conditions and timber supply, there is increasing demand for regional and national data layers depicting forest cover. These data layers must permit small area estimates of forest area and, most importantly, provide associated error estimates. This paper presents a model-based approach for coupling mid-resolution satellite imagery with plot-based forest inventory data to produce estimates of probability of forest and associated error at the pixel-level. The proposed Bayesian hierarchical model provides access to each pixel's posterior predictive distribution allowing for a highly flexible analysis of pixel and multi-pixel areas of interest. The paper presents a trial using multiple dates of Landsat imagery and USDA Forest Service Forest Inventory and Analysis plot data. The results describe the spatial dependence structure within the trial site, provide pixel and multi-pixel summaries of probability of forest land use, and explore discretization schemes of the posterior predictive distributions to forest and non-forest classes. Model prediction results of a holdout set analysis suggest the proposed model provides high classification accuracy, 88%, for the trial site.

A. O. Finley (✉)
Department of Forestry and Department of Geography, Michigan State University,
126 Natural Resources Building, East Lansing, MI 48824, USA
e-mail: finleya@msu.edu

S. Banerjee
Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building,
MMC 303, 420 Delaware Street S.E., Minneapolis, MN 5545, USA
e-mail: Sudiptob@biostat.umn.edu

R. E. McRoberts
Forest Inventory and Analysis, Northern Research Station, USDA Forest Service,
1992 Folwell Ave. St. Paul, MN 55108, USA
e-mail: rmcroberts@fs.fed.us

Keywords Bayesian inference · Forest inventory · Logistic model · Markov Chain · Monte Carlo · Metropolis–Hastings · Spatial process models

1 Introduction

In areas such as land use change monitoring, carbon budgeting, and ecological and timber supply forecasting, there is an increasing demand for spatially explicit estimates of forest area. In response to this need, many countries have established national inventories designed to provide agencies and researchers with these base data (e.g., National Inventory of Landscapes in Sweden, National Finnish Forest Inventory, Canadian National Forest Inventory, and the National Forest Inventory of Switzerland).

Many of these large forest inventory programs couple inventory data collected from field plots and ancillary data, such as satellite imagery, to form land use strata. These strata are then used in design-based inference of forest attributes. Satellite imagery has provided a useful and cost effective source for deriving the data layers required for stratified estimation (McRoberts et al. 2002). These stratified estimation techniques can produce satisfactory estimates and precision for medium to large geographic areas, but they typically fail to satisfy precision expectations for small areas. Obtaining forest area estimates for small areas requires more spatially intensive sampling designs, more and different kinds of ancillary data, and/or methods that extract more information from inexpensive sources of ancillary data. The increased costs associated with more intense sampling and a larger suite of ancillary data often precludes these approaches. Therefore, approaches to make better use of common and affordable satellite imagery merit consideration.

This paper presents a model-based approach that couples field inventory data from the Forest Inventory and Analysis (FIA) program of the USDA Forest Service with mid-resolution satellite imagery to predict pixel-level forest probability with associated error estimates. The Bayesian hierarchical model presented provides access to each pixel's full predictive distribution from which we calculate the desired inferential statistics. Further, when combined with an appropriate area estimator, these individual pixel estimates can provide area and error estimates for arbitrary areas of interest (AOI).

The strength of the approach we describe is accessibility to pixel and multi-pixel posterior predictive distributions. Many classifiers commonly used in forest area mapping can only offer estimates of overall classification accuracy. For example, the popular k-nearest neighbor classifier might use a leave-one-out cross-validation to provide a measure of expected classification accuracy for the entire mapping extent (Franco-Lopez et al. 2001; Tomppo 1991) but cannot provide spatially explicit estimates of precision for small areas within the mapped extent. Our proposed framework precisely addresses this latter issue.

This paper is organized as follows. Section 2 presents a trial data set comprised of FIA field inventory plots and satellite imagery from the Landsat sensors. Section 3 reviews the basic logistic model, followed by a description of a Bayesian hierarchical model for incorporating spatial structure within the context of the inventory data.

Parameter estimation and prediction are then described. Trial results are in Sect. 4. Discussion and concluding remarks are given in Sect. 5.

2 Data and trials

2.1 Forest inventory plot data

The FIA program of the USDA Forest Service has established field plot centers in permanent locations using a sampling design that is assumed to produce a systematic equal-probability sample with a random spatial component (Bechtold and Patterson 2005). Locations of forested plots are determined using global positioning system (GPS) receivers, and locations of non-forested plots are determined using aerial imagery and digitization methods. Each plot consists of four 7.31 m radius circular subplots. The subplots are configured as a central subplot and three peripheral subplots with centers located 36.58 m and azimuths of 0°, 120°, and 240° from the center of the central subplot. The distance between the peripheral subplots is 63.00 m.

At each subplot, field crews record the proportion of area that satisfy specific ground land use conditions. Subplot estimates of proportion forest area are obtained by collapsing ground land use conditions into forest and non-forest classes consistent with the FIA definition of forest land (Bechtold and Patterson 2005).

2.2 Satellite imagery

Landsat imagery for one Indiana scene, row 21 of path 23, was obtained from the MultiResolution Characterization 2001 land cover mapping project (Homer et al. 2004) of the U.S. Geological Survey. Three dates of imagery were acquired: April 30 2001, July 8 2001, and October 31 2002, corresponding to early and peak vegetation green-up and senescence. The April and July images are from the Landsat 5 TM sensor and the October image is from the Landsat 7 ETM+sensor. All images were georectified to a common base layer, each with a root mean square error of less than 30 m. In the rectification process, images were resampled to a 30×30 m spatial resolution using the cubic convolution algorithm (Campbell 1996).

Each date of imagery was tasseled cap transformed into its brightness, greenness, and wetness components then scaled into 8-bit [0,255] (Kauth and Thomas 1976). These nine spectral variables are reference as the image month concatenated with tassel cap (TC), brightness (1), greenness (2), and wetness (3): AprilTC1, AprilTC2, AprilTC3, JulyTC1, JulyTC2, JulyTC3, and OctTC1, OctTC2, OctTC3.

2.3 Combining FIA data and satellite imagery

The spatial configuration of the FIA subplots with centers separated by 36.58 m and the 30×30 m spatial resolution of the imagery allows each subplot centroid to be uniquely associated with the pixel with which it is spatially aligned. Our analysis relates land use observed at the FIA subplots with the spectral values from the Landsat

sensors. Therefore, several source of error must be acknowledged. First, the subplot only represents approximately 19% of the pixel area. As a result, the subplot might not adequately represent the proportion of forest area for the entire pixel. Second, pixel values were assigned based on the subplot center coordinate; therefore, it is possible that a subplot area might cover portions of four or fewer pixels. Third, GPS and image registration error might cause a subplot to be associated with the incorrect pixel, resulting in the subplot observation being erroneously matched with a vector of spectral values from a non-forested pixel, and vice versa. These sources of error obscure the relationship between forest probability and spectral characteristics, increase the uncertainty in model parameter estimates, and ultimately increase the variance in the pixel-level predictive distribution.

2.4 Trial

To demonstrate the proposed model, a trial site that covers a mix of forest and non-forest land use was selected within the Landsat scene described above. Several land use conditions can exist within a given subplot (e.g., water, forest, non-forest); however, for this analysis only single condition forest and non-forest land use subplots were considered. For the analysis, 500 subplots closest to the trial site centroid were selected. From this sample, a holdout set of 25 plots (100 subplots) were randomly selected and used for model validation. The remaining 400 subplots were used for model construction. Within this set, 181 subplots were designated as forest and 219 were non-forest. All subplots were observed between the beginning of 1999 and the end of 2003.

Within the model construction set, the sampling intensity is about one plot every 33 km^2. The maximum distance between any two plots is 63.85 km. The maximum, minimum, and mean distance between any two nearest neighbor plots is 6,303.72, 1,142.85, and 3,657.55 m, respectively. The maximum, minimum, and mean distance between any plot and its second nearest neighbor plot is 8,351.15, 2,552.71, and 4,954.41 m, respectively.

For the trial a 32 km radius circle, centered on the trial site's center, was clipped from the nine variable image stack and used for mapping probability of forest and associated error estimates. Further, within this image circle, 15 20×20 pixel (36 ha) AOIs were selected to illustrate estimation of multi-pixel forest proportion. Based only on visual interpretation of the raw Landsat images, the AOIs were selected to represent areas with high, moderate, and low proportion of forested pixels.

3 Statistical modelling

3.1 Non-spatial logistic model

We first outline a basic logistic model that can be used for modelling the forestation. Suppose we have $i = 1, \ldots, n$ subplots. We set y_i as the binary variable designating this classification with $y_i = 1$ denoting that subplot i is forested and $y_i = 0$ otherwise. Conditional upon the set of predictor variables (spectral characteristics

for us), say $\mathbf{x}_i$ for subplot i, we assume that the y_i's follow a Bernoulli distribution, $y_i \overset{i.i.d}{\sim} Ber(p(\mathbf{x}_i))$ with $P(y_i = 1 \mid \mathbf{x}_i) = p(\mathbf{x}_i)$. The association between the response data vector $\mathbf{y} = (y_1, \ldots, y_n)$, and the $n \times m$ matrix of m spectral predictor variables $X = [\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T]$, where each $\boldsymbol{x}_i^T$ is the $1 \times m$ vector of spectral characteristics for the i-th point, is modelled through a logistic link regression,

$$p(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\theta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\theta})}, \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ is the vector of parameters to be estimated.

Letting $Data$ denote all the available information, say $\mathbf{y}$, X above, the likelihood function for the data given the above model is,

$$L(\boldsymbol{\theta};\ Data) = \prod_{i=1}^{n} p(\mathbf{x}_i)^{y_i} (1 - p(\mathbf{x}_i))^{1-y_i} = \prod_{i=1}^{n} \frac{\exp\left(\mathbf{x}_i^T \boldsymbol{\theta}\right)^{y_i}}{1 + \exp\left(\mathbf{x}_i^T \boldsymbol{\theta}\right)}. \tag{2}$$

This yields the corresponding log-likelihood function as

$$\ln\left(L(\boldsymbol{\theta};\ Data)\right) = \sum_{i=1}^{n} y_i \left(\mathbf{x}_i^T \boldsymbol{\theta}\right) - \sum_{i=1}^{n} ln(1 + \exp(\mathbf{x}_i^T \boldsymbol{\theta})). \tag{3}$$

Typically, from (3) iterative methods are used to obtain the maximum likelihood estimates of the parameters, $\hat{\boldsymbol{\theta}}$. These, however, rely upon asymptotic (for large samples) distributional assumptions that are rarely verifiable in practice (Ferguson 1996). Alternatively, we adopt a Bayesian paradigm (e.g., Gelman et al. 2004) that enables direct probabilistic inference for all the model parameters by first specifying prior distributions for them and subsequently using the likelihood in (2) to obtain the posterior distribution. In practice, therefore, if $p(\boldsymbol{\theta})$ is the prior distribution for $\boldsymbol{\theta}$, the posterior distribution of $\boldsymbol{\theta}$ is given by:

$$p(\boldsymbol{\theta} \mid Data) \propto p(\boldsymbol{\theta}) L(\boldsymbol{\theta};\ Data).$$

Markov chain Monte Carlo (MCMC) integration methods (e.g., Gelman et al. 2004) provide samples from the full posterior distribution of $\boldsymbol{\theta}$ that can subsequently be used for inference.

3.2 Logistic model with spatial random effects

The unexplained residual uncertainty associated with the mean function in (1) does not accommodate spatial correlation among subplot observations, which can impair the precision of predictions. A multi-stage hierarchical model allows us to explicitly incorporate spatial structure into the basic model. Indeed, for data sets revealing complex variability patterns, building models hierarchically allow enormous richness in

capturing variability by incorporating estimable parameters that should explain different sources of variation. Partitioning sources of variance through a hierarchical specification often results in much better model fit, compared to fits obtained with simpler (single stage) models.

The first stage of the hierarchical model assumes that the observed responses over the locations are conditionally independent given the spatial effects and adds spatially correlated random effects to the mean structure in (1). The second stage specifically models the nature of association between these random effects. Finally, a full hierarchical specification is achieved by using prior distributions for the model parameters. This is essentially the paradigm of Bayesian modelling (Gelman et al. 2004; Banerjee et al. 2004).

Specifically for our setting, suppose the subplots are spatially referenced (e.g., Easting-Northing or some other coordinate system) as $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$. Then we can envision the response as $y(\mathbf{s}_i) = 1$ or 0 depending upon whether the subplot is forested. Within the augmented model, the probability that $y(\mathbf{s}_i) = 1$ depends upon spatially-referenced predictor variables, $\mathbf{x}(\mathbf{s}_i)$ for subplot $\mathbf{s}_i$, the regression slope parameters $\boldsymbol{\theta}$, and the location-specific random effects $w(\mathbf{s}_i)$ to yield:

$$p(\mathbf{s}_i) = \frac{\exp\left(\mathbf{x}(\mathbf{s}_i)^T\boldsymbol{\theta} + w(\mathbf{s}_i)\right)}{1 + \exp\left(\mathbf{x}(\mathbf{s}_i)^T\boldsymbol{\theta} + w(\mathbf{s}_i)\right)}. \tag{4}$$

In the present context, $\mathbf{s} \in D$, which defines the surface of interest within $\Re^2$.

The second stage of the hierarchical model specifies the association in the random effects. A popular specification for the random effect is the Gaussian Process, denoted by $w(\mathbf{s}) \sim GP(\mu(\mathbf{s}), K(\cdot))$ where $\mu(\mathbf{s})$ is the process mean (or trend surface) and $K(\cdot)$ is a positive definite covariance function. Gaussian processes are extremely popular in modelling spatial variation, due to their ability to directly model spatial correlation. More extensive treatments can be found, for example, in Cressie (1993), Chilés and Delfiner (1999) and Banerjee et al. (2004).

We assume $w(\mathbf{s}) \sim GP(0, K(\phi))$, where $K(\mathbf{s} - \mathbf{s}'; \phi) = \sigma^2\rho(\mathbf{s} - \mathbf{s}'; \phi)$, which means that for any arbitrary collection of sites, say $\mathcal{S}$, the corresponding process realization $\mathbf{w} = (w(\mathbf{s}_i))_{i=1}^N$ follows a multivariate Gaussian distribution $MVN(\mathbf{0}, \sigma^2 R(\phi))$, where $R(\phi) = [\rho(\mathbf{s}_i - \mathbf{s}_j; \phi)]_{i,j=1}^N$ is the $N \times N$ spatial correlation matrix. Here the strength of spatial association is captured through a spatial correlation function, $\rho(\mathbf{s}-\mathbf{s}'; \phi)$. These functions are also known as *positive definite* functions as they must ensure that the matrix $R(\phi)$ is positive definite—for any collection of sites in $\mathcal{S}$. Bochner's theorem (see, e.g., Banerjee et al. 2004) characterizes the characteristic functions of symmetric random variables as an exhaustive class of real-valued positive definite functions. The exponential correlation function, $\rho(\mathbf{s} - \mathbf{s}'; \phi) = \exp(-\phi\|\mathbf{s} - \mathbf{s}'\|)$, is among the more popular choices for its easier interpretability and is used in the current analysis. The parameters associated with the exponential function and subsequent covariance matrix, $\sigma^2 R(\phi)$, are the spatial decay parameter ϕ and the spatial effect variance σ^2. We describe the effective range d_0 of the spatial process by solving $\exp(-\phi d_0) = 0.05$ (i.e., $d_0 \approx 3/\phi$).

Finally, prior probability distributions are assigned to the model parameters that, together with the data likelihood, yields a fully specified Bayesian hierarchical model. We discuss this in detail in the following section.

3.3 The priors and likelihood

With the addition of the random effects, the parameter set is $\mathbf{\Omega} = (\boldsymbol{\theta}, \sigma^2, \phi, \mathbf{w})$ with length $m + 2 + n$. The Bayesian hierarchical specification assigns prior distributions to the parameters. Choice of priors can play an important role in the efficiency of the algorithm. In the subsequent analysis, a non-informative (or uninformative) flat prior is assigned to the fixed effects $\boldsymbol{\theta}$ (i.e., $p(\boldsymbol{\theta}) \propto 1$). The spatial effect variance parameter, σ^2, is assumed to follow an inverse-Gamma distribution, $\sigma^2 \sim IG(a_\sigma, b_\sigma)$. Fixing $a_\sigma = 2$, this distribution has an infinite variance, and its mean is b_σ (as defined in Appendix A of Gelman et al. 2004). This family is a widely used specification for variance parameters, as it allows the modeler to center the prior on a reasonable belief while maintaining a large prior variance. The large variance of this prior should allow the data to overwhelm the prior beliefs and dominate the inference. The spatial decay parameter received a Uniform, $\phi \sim U(a_\phi, b_\phi)$. Here, we emphasize the need for a fairly informative proper prior for the ϕ parameter that will ensure proper and well-identified posteriors and stabler convergence of algorithm (e.g., Berger et al. 2001). Such information is usually based upon the configuration of the locations and, in particular, the distances between the sites. For instance, it is plausible that the spatial range would not exceed the maximum intersite distance and would always exceed a certain minimal distance. The parameters a_ϕ and b_ϕ can be determined based upon these specifications.

The posterior distribution for $\mathbf{\Omega}$ becomes

$$P(\mathbf{\Omega} \mid Data) \propto P(\phi)P(\boldsymbol{\theta})P(\sigma^2)P(\mathbf{w} \mid \sigma^2, \phi) \times L(\mathbf{\Omega};\ Data), \tag{5}$$

where $L(\mathbf{\Omega};\ Data)$ is the data likelihood depending upon the generic parameter set $\mathbf{\Omega}$. In our current setting, the data likelihood constitutes the first stage of our hierarchical model. It can be expressed more specifically as $L(\mathbf{w}, \boldsymbol{\theta}; \mathbf{y}, X)$, where $\mathbf{y} = (y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n))^T$ is the $n \times 1$ response vector observed over locations in $\mathcal{S}$, and $X = [\mathbf{x}^T(\mathbf{s}_i)]_{i=1}^n$ is the $n \times p$ matrix of independent predictors, which we often suppress in the subsequent equations. Thus, $L(\mathbf{w}, \boldsymbol{\theta};\ \mathbf{y}, X)$ is modified from (2) to incorporate the spatial effects

$$\begin{aligned} L(\mathbf{w};\ Data) = \prod_{i=1}^{n} P(y(\mathbf{s}_i)\ &= 1|\mathbf{x}(\mathbf{s}_i), \boldsymbol{\theta}, \mathbf{w}(\mathbf{s}_i))^{y(\mathbf{s}_i)}(1 - P(y(\mathbf{s}_i) \\ &= 1|\mathbf{x}(\mathbf{s}_i), \boldsymbol{\theta}, \mathbf{w}(\mathbf{s}_i)))^{1-y(\mathbf{s}_i)}, \end{aligned} \tag{6}$$

with $P(y(\mathbf{s}_i) = 1) = logit^{-1}(\mathbf{x}^T(\mathbf{s}_i)\boldsymbol{\theta} + w(\mathbf{s}_i))$. The Gaussian Process specification implies that the $P(\mathbf{w} \mid \boldsymbol{\theta})$ is a multivariate normal $MVN(\mathbf{0}, \sigma^2 R(\phi))$. The log-posterior is now written as

$$\ln\left(p(\boldsymbol{\Omega}\,|\,\mathbf{y})\right) \propto -\left(a_\sigma + 1 + \frac{n}{2}\right)\ln\left(\sigma^2\right) - \frac{b_\sigma}{\sigma^2} - \frac{1}{2\sigma^2}\mathbf{w}^T R^{-1}\mathbf{w} - \frac{1}{2}\ln\left(|R|\right) + \sum_{i=1}^{n} y(\mathbf{s}_i)\left(\mathbf{x}(\mathbf{s}_i)^T\boldsymbol{\theta} + w(\mathbf{s}_i)\right) - \sum_{i=1}^{n}\ln\left(1+\exp(\mathbf{x}(\mathbf{s}_i)^T\boldsymbol{\theta} + w(\mathbf{s}_i)\right). \tag{7}$$

In the numerical implementation, the prior on ϕ is treated a bit differently. As stated above, its prior is Uniform with the condition,

$$p(\phi) = \begin{cases} \frac{1}{b_\phi - a_\phi} & \text{if } \phi \in (a_\phi, b_\phi), \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

In principle, this condition is problematic when the posterior is log-transformed (i.e., $\ln(0) = -\infty$); however, this is easily treated in the sampling approach described in the following section.

3.4 Posterior sampling

The Metropolis–Hastings algorithm was used to generate the marginal posterior distribution for each parameter in $\boldsymbol{\Omega}$. Initially, candidate values for the parameters were drawn as a single block from a multivariate normal density. In an attempt to maintain a ∼23% acceptance rate (Gelman et al. 2004), we adjusted the diagonal elements (i.e., the tuning values) of the multivariate normal Σ matrix. However, we experienced difficulty in achieving a reliable acceptance rate that would indicate sufficient mixing. In fact the acceptance rate in our initial trials was typically ≤1%, while a healthy rate should hover around 23% (Gelman et al. 2004). This is not unusual with joint-Metropolis updates: although they are simpler to implement (entailing a single likelihood evaluation for each iteration), with high-dimensional non-Gaussian likelihoods and with less informative priors, these single block-updates may take a very long to converge (i.e., to find their stationary distribution). Therefore we split $\boldsymbol{\Omega}$ into its components, and drew candidate values for $\boldsymbol{\theta}$, σ^2, ϕ, and $\boldsymbol{w}$ separately. This required four sequential Metropolis–Hastings steps, where $\boldsymbol{\theta}$ and $\boldsymbol{w}$ were block updated. In this scheme, we monitored four separate acceptance rates, and generally found much better mixing; this was further improved by specifying the covariance structure among the $\boldsymbol{\theta}$ as the dispersion of a multivariate normal proposal for $\boldsymbol{\theta}$.

As noted in the previous section, the Uniform prior on ϕ required that it be treated differently than the other parameters. Specifically, each candidate value of ϕ drawn from the normal proposal density was applied to the conditional statement (8). If the candidate ϕ passed (8), it proceeded through the Metropolis–Hastings iteration; otherwise, the candidate was discarded and subsequent candidates were drawn until the condition was satisfied.

3.5 Convergence diagnostics

The algorithms were written in C++ and used the Intel® Math Kernel Library BLAS and LAPACK routines. Posterior samples were formatted to be read by the Convergence Diagnostics and Output Analysis (CODA) available in R (see http://www.R-project.org).

Multiple independent chains were run for the trial. Each chain was given a unique seed for the program's random number generator and starting values for the parameters were
dispersed across each parameters' feasible range. The chains were graphed using appropriate trace plots (available in CODA) and the Gelman–Rubin diagnostics were computed.

3.6 Model comparison

For the trial, the simple logistic model is compared to the logistic model with spatial random effects using the *deviance information criterion* (DIC), proposed by Spiegelhalter et al. (2002). DIC is similar to Akaike Information Criterion (AIC) in that it penalizes larger models; however, DIC is more suited to hierarchical models as it estimates the complexity, unlike AIC which assumes the penalty is known. This criterion is based on the posterior distribution of the *deviance* statistic,

$$D(\mathbf{\Omega}) = -2 \ln L(\mathbf{\Omega};\ Data) + 2 \ln h(Data), \tag{9}$$

where $L(\mathbf{\Omega};\ Data)$ is the data likelihood given model parameters $\mathbf{\Omega}$ and $h(Data)$ is some standardizing function of the data alone (thus can be dropped with no impact to model selection). In this approach, the *fit* of the model is summarized by the posterior expectation of the deviance, $\overline{D} = E_{\mathbf{\Omega}\,|\,Data}[D]$, and the *complexity* of a model is captured by the effective number of parameters, p_D. Spiegelhalter et al. (2002) show that a reasonable definition of p_D is

$$p_D = E_{\mathbf{\Omega}\,|\,Data}[D] - D(E_{\mathbf{\Omega}\,|\,Data}[\mathbf{\Omega}]) = \overline{D} - D(\overline{\mathbf{\Omega}}), \tag{10}$$

where $\overline{\mathbf{\Omega}}$ is the mean of the parameters' samples. Typically, this effective parameter total, p_D, will be less than the actual total number of parameters in the model due to collinearity among the variables and borrowing of strength across random effects. The DIC is then defined analogously to the AIC as the expected deviance plus the effective number of parameters,

$$DIC = \overline{D} + p_D. \tag{11}$$

Because small values of $\overline{D}$ suggest good fit and small values of p_D indicate a parsimonious model, the preferred models will have lower DIC. As with AIC and other penalized likelihood criteria, DIC is not a metric for identifying the 'correct' model, but merely a metric to compare a collection of alternative formulations (all of which

might be incorrect). As alluded to above, DIC is scale-free; the choice of standardizing function $h(Data)$ in (9) is arbitrary. Thus, values of DIC have no intrinsic meaning, only differences in DIC across models are meaningful.

3.7 Prediction

Predictions can be made once the samples $\{\mathbf{\Omega}^{(k)}\}_{k=1}^{N}$ are obtained from the posterior distribution $P(\mathbf{\Omega} \mid Data)$. The posterior *predictive* distribution we seek is

$$P(y(\mathbf{s}_0) = 1 \mid \mathbf{y}, \mathbf{x}, \mathbf{x}(\mathbf{s}_0)) = \int P(y(\mathbf{s}_0) = 1 \mid \mathbf{\Omega}, \mathbf{y}, \mathbf{x}(\mathbf{s}_0)) P(\mathbf{\Omega} \mid \mathbf{y}, \mathbf{x}) d\mathbf{\Omega}. \quad (12)$$

where $\mathbf{s}_0$ denotes the location for which the vector $\mathbf{x}(\mathbf{s}_0)$ is known and we wish to predict y. Samples from (12) are obtained by *composition* sampling: for each $\mathbf{\Omega}^{(k)}$ from the posterior sample we simply compute $P(y(\mathbf{s}_0) = 1 \mid \mathbf{\Omega}^{(k)}, \mathbf{y}, \mathbf{x}(\mathbf{s}_0))$ for $k = 1, \ldots, N$. Programmatically, we first generate a vector of N samples from $\mathbf{s}_0$'s location effect with each element defined by a draw from a normal distribution with mean

$$\boldsymbol{\varphi}_0^{(k)T} R^{-1}(\phi^{(k)}) \mathbf{w}^{(k)} \quad (13)$$

and variance

$$\sigma^{2(k)} \left[1 - \boldsymbol{\varphi}_0^{(k)T} R^{-1}(\phi^{(k)}) \boldsymbol{\varphi}_0^{(k)} \right], \quad (14)$$

where $\boldsymbol{\varphi}_0^{(k)}$ is the $n \times 1$ vector with i-th element given by $\boldsymbol{\varphi}_{0i}^{(k)} = exp(-\phi^{(k)} \| \mathbf{s}_0 - \mathbf{s}_i \|)$. Then, a vector of probabilities is generated with each element defined by (4) replacing $\mathbf{x}(s_i)$ and $w(\mathbf{s}_i)$ with $\mathbf{x}(s_0)$ and $w(\mathbf{s}_0)$. The resulting sample is precisely a sample from the desired predictive distribution in (12).

A forest probability map can be created using the posterior mean or median (or, for that matter, any other quantile) by simply carrying out the above predictive sampling over a grid of sites. Creating the associated uncertainty map for these predictions is just as simple. For the grid of sites, compute the uncertainty summary (standard deviation or range) from the predictive sample. We point out that the standard deviations computed from MCMC output are biased as these samples are correlated (Gelman et al. 2004). However, this bias becomes negligible as the size of the MCMC sample becomes large. This sample, being in our control (subject to computational limits), is usually taken large enough and this issue is not serious.

3.8 Estimating multiple pixel AOI

Once complete posterior distributions are obtained for the pixel-level forest probabilities, interest often turns to obtaining forest area for multi-pixel AOIs. To be precise, suppose we are interested in a region composed of N_A pixels, say $A = \cup_{i=1}^{N_A} \{\mathbf{s}_i\}$

(perhaps after suitable relabelling of the $\mathbf{s}_i$'s). An estimate of the fraction of the forest area in A is given in terms of the corresponding probabilities at the pixel level by

$$F_A = \frac{1}{N_A} \sum_{i=1}^{N_A} P(Y(\mathbf{s}_i) = 1). \tag{15}$$

Hence, samples $\{F_A^{(k)}\}_{k=1}^{N_{MCMC}}$ from the posterior distribution $p(F_A|Data)$ are immediately obtained from $\{P^{(k)}(Y(\mathbf{s}_i)|Data)\}_{k=1}^{N}$ using (15), once the latter are obtained using the methods described in the preceding section. Note that any other functional, such as the total area inside A under forests $\tilde{F}_A = |A|F_A$, where $|A|$ denotes the area of the region A, is also immediately accessible to posterior inference.

4 Trial results

4.1 Priors and model convergence

As described in Sect. 3.3, a flat prior was assigned to the $\boldsymbol{\theta}$. The variance term σ^2 received an inverse-Gamma prior $IG(2, b_\sigma)$ with infinite variance and mean b_σ. Fitting a single stage logistic regression model with random effects and setting b_σ equal to the maximum likelihood estimate usually yields a reasonable centered, yet vague, prior. The spatial decay parameter received a Uniform, $\phi \sim U(a_\phi, b_\phi)$, where $b_\phi = 3$ and $a_\phi = 2\text{e}{-4}$, which sets an effective spatial range of 1–15,000 m: one that is about a quarter of the trial site's diameter. Again, these priors were robust for our analysis and we witnessed substantial posterior learning from the data.

For the trial, five Metropolis–Hastings chains were run. Acceptance rates for $\boldsymbol{\theta}$, σ^2, ϕ, and $\mathbf{w}$ were between 15% and 30% for all chains. Each chain was run for 150,000 iterations. The CODA package was used to diagnose convergence by monitoring mixing using Gelman–Rubin diagnostics and autocorrelations (e.g., Gelman et al. 2004, Section 11.6). These revealed sufficient mixing of the chains after 30,000 iterations. Therefore, the first 30,000 iterations were discarded as burn-in and the remaining posterior samples were thinned for every 10 iterations and then combined to yield 60,000 samples ($5 \times 120,000/10$) for parameter estimation and prediction.

Following the criteria discussed in Sect. 3.6, the values $\overline{D}$, p_D, and DIC for the non-spatial logistic model were 112.60, 10.57, and 123.17, respectively. For the spatial model these values were 89.96, 14.03, and 103.97, respectively. Despite the greater number of parameters, the lower DIC values support the spatial model over the non-spatial model. The discrepancy between the effective number of parameters p_D and the true number, suggests that within the spatial model there is substantial shrinkage toward the overall mean of the spatial process. We choose to continue the analysis with the spatial model, based on the DIC criteria and the understanding that the spatial dependence structure among observations is required for calculating unbiased variance estimates on both model parameters and subsequent predictions.

4.2 Parameter estimates, model validation, and prediction

Parameter estimates and predictions for the trail were based on 400 subplots. Table 1 offers the 2.5, 50, and 97.5 percentiles for the model parameters. The credible intervals suggest that the intercept, all spring tasseled cap variables, and fall brightness and wetness variables significantly contribute to the model fit. The median of the σ^2 and ϕ is 1.36 and 0.00182, respectively. The distribution of the spatial decay parameter ϕ, describes a process with spatial dependence within distances of 1,644.19 m. The estimate of ϕ indicates strong within plot dependence (i.e., among subplots within a plot). However, the large credible interval about the point estimate suggests there is substantial variability in between plot dependence; specifically, the credible interval suggests there might be dependence extending to the first an perhaps second nearest neighbor plot.

The model validation used a holdout set of 25 plots. Of these 100 subplots, 88 were correctly classified when using a cut-point of 0.5. Table 2 provides the prediction summary for only the center subplots of the 25 holdout plots. Although the majority of the predictions correctly place their entire distribution above or below the 0.5 cut-point, several predictive distributions straddle the cut-point (e.g., subplots 2, 7, 9, 11, 20, 24, and 25 in Table 2) and one entirely misclassified the subplot (i.e., subplot 21 in Table 2).

Figures 1 and 2 provide the probability map and associated error for the trial site. The median serves as the pixel-level point prediction (Fig. 1) and the error is represented by the range between the 0.25 and 0.975 quantiles (Fig. 2). Based on the shape and sizes of the low probability forest patches it appears that but for a relatively contiguous forest range in the northwest, the trial site is predominately under agricultural land use.

Table 1 Parameter estimates the trial logistic model with spatial random effects

Parameters	Estimates: 50% (2.5%, 97.5%)
Intercept (θ_0)	82.39 (49.56, 120.46)
AprilTC1 (θ_1)	−0.27 (−0.45, −0.11)
AprilTC2 (θ_2)	0.17 (0.07, 0.29)
AprilTC3 (θ_3)	−0.24 (−0.43, −0.08)
JulyTC1 (θ_4)	−0.04 (−0.25, 0.17)
JulyTC2 (θ_5)	0.09 (−0.01, 0.19)
JulyTC3 (θ_6)	0.01 (−0.15, 0.16)
OctTC1 (θ_7)	−0.43 (−0.68, −0.22)
OctTC2 (θ_8)	−0.03 (−0.19, 0.14)
OctTC3 (θ_9)	−0.26 (−0.46, −0.07)
σ^2	1.358 (0.39, 2.42)
ϕ	0.00182 (0.00065, 0.0032)
log(0.05)/ϕ (m)	1644.19 (932.33, 4606.7)

Table 2 Trial predicted probability of forest for center subplots of the holdout plot set. Observed value is FIA recorded forest (1) and non-forest (0)

Plot	Observed value	Estimates: 50% (2.5%, 97.5%)
1	0	0.000 (0.000, 0.001)
2	1	0.428 (0.043, 0.924)
3	0	0.000 (0.000, 0.005)
4	1	0.998 (0.951, 1.000)
5	1	0.940 (0.508, 0.996)
6	1	0.981 (0.729, 0.999)
7	1	0.900 (0.355, 0.993)
8	0	0.000 (0.000, 0.000)
9	0	0.139 (0.007, 0.769)
10	0	0.000 (0.000, 0.001)
11	0	0.066 (0.000, 0.890)
12	1	0.995 (0.860, 1.000)
13	1	0.993 (0.889, 1.000)
14	0	0.001 (0.000, 0.044)
15	1	0.987 (0.840, 0.999)
16	0	0.000 (0.000, 0.004)
17	0	0.027 (0.000, 0.531)
18	0	0.000 (0.000, 0.013)
19	0	0.000 (0.000, 0.007)
20	0	0.057 (0.002, 0.627)
21	0	0.966 (0.589, 0.998)
22	1	0.990 (0.873, 0.999)
23	1	0.000 (0.000, 0.000)
24	0	0.621 (0.065, 0.974)
25	0	0.810 (0.206, 0.987)

4.3 Prediction for small AOIs

To illustrate small area prediction we selected 15 AOIs within the trial site. The AOIs were selected prior to the analysis and based only on visual interpretation of the raw Landsat imagery. These fixed area (36 ha) AOIs were chosen to represent high, moderate, and low proportion of forest. Access to the posterior predictive distribution of each pixel allows for straightforward generalization to the posterior predictive distribution of arbitrary AOIs. For each AOI, the mean of its posterior predictive distribution was calculated by (15). For the trial site, the mean of the predictive distributions for high proportion forest AOIs ranged from 0.89 to 0.99, moderate ranged from 0.32 to 0.71, and low ranged from 0 to 0.13.

Point estimates of the first and perhaps second order statistics of probability forest in a given AOI are useful; however, access to the full posterior predictive distribution

Fig. 1 Median value of pixel-specific posterior predictive distribution for probability of forest across the trial site. The maximum probability in the image is 1.00 (*black*) and the minimum probability is 0.00 (*white*). The 0.25, 0.50, and 0.975 quantiles of pixel values across the image are 0.00, 0.81, and 0.99, respectively

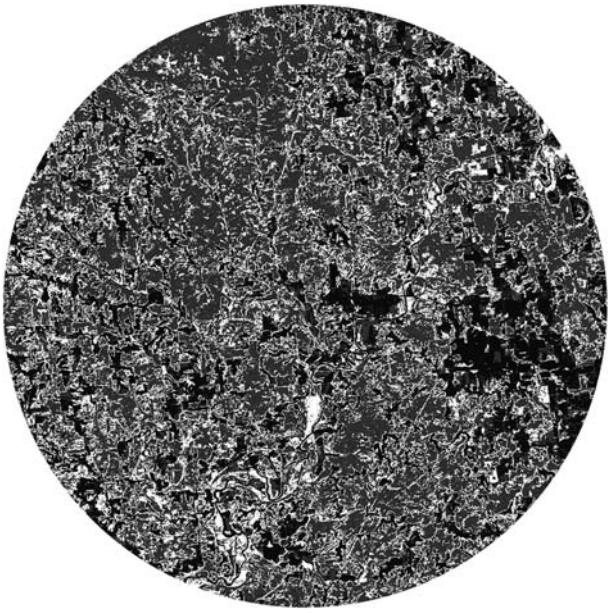

Fig. 2 Range between the 0.25 and 0.975 quantiles of pixel-specific posterior predictive distribution for probability of forest across the trial site. This is a measure of precision for the estimates in Fig. 1. The maximum range in the image is 0.99 (*white*) and the minimum range is 0.00 (*black*). The 0.25, 0.50, and 0.975 quantiles of pixel values across the image are 0.02, 0.13, and 0.99, respectively

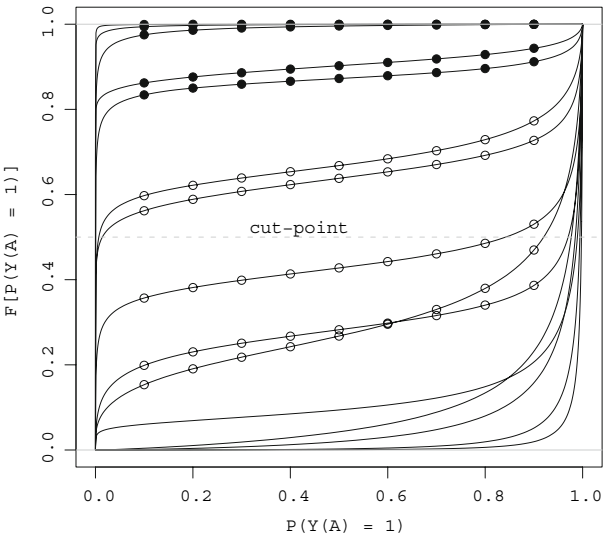

Fig. 3 CDF of the posterior predictive distribution for probability of forest within AOIs across the trial site. AOIs of high, moderate, and low forest area were selected based on (pre-analysis) visual interpretation of raw Landsat imagery. Solid circle markers denote AOIs thought to have low forest area. Open circle markers denote AOIs thought to have moderate forest area. Lines with no point markers represent AOIs thought to have high forest area. All AOIs are 20×20 pixels and represent 36.0 ha in area

can provide greater insight and flexability of analysis. Figure 3 offers the empirical cumulative distribution function (CDF) and potential 0.5 cut-point for probability of forest within AOIs. These CDFs suggest that forest and non-forest dominated AOIs have very defined probability curves, whereas AOIs with moderate forest area present a bimodal distribution.

5 Discussion

5.1 Trial results

The Bayesian hierarchical model presented provides access to each pixel's full predictive distribution from which the desired inferential statistics are calculated. At the first stage of the hierarchy, the mean function of the logistic link regression is augmented by random spatial effects. The second stage defines the Gaussian Process from which these random effects arise. Based on our definition of the spatial process, the analysis suggests that dependence typically extends beyond the within plot subplots to the first and perhaps the second nearest neighbor plots. Further, the credible intervals identified several covariates that contribute significantly to forest/non-forest discrimination.

Once model parameters were estimated, we used composition sampling to detail each "new" pixel's posterior predictive distribution. Based on these distributions, we mapped the median and range between the 0.25 and 0.975 percentiles. However, any percentile or function of the predictive distribution can be mapped. Beyond describing

the uncertainty in probability of forest estimates, error maps can reveal missing covariates and paucity of model "training" observations for certain regions or land use classes. Within Fig. 2, there seems to be high prediction error associated with areas adjacent to rivers. This suggests that spectral signatures associated with the land use of alluvial plains might not be adequately represented in the observation set used for parameter estimation, which might lead us to gather additional observations within this land use class.

Further, the prediction error maps reveal low precision at the boundary of well established forest and non-forest areas or patches. Boundary pixels are often referred to as mixed pixels in the remote sensing literature, and as the prediction error suggests, mixed pixels are difficult to classify because of the sub-pixel mixture of land use classes and hence spectral characteristics. See Campbell (1996, p. 378), for a general discussion on the challenges associated with mixed pixel classification.

The sampling design used by FIA is characterized by large distances among plots (e.g., thousands of meters) and relatively small within plots distances (i.e., less than 63 m among subplots). In this trial, the average distance between any two plots is 3,657.55 and the average distance between any plot and its second nearest neighbor plot is 4,954.41. This disparity in the concentration of observations will decrease the precision in parameter estimates, specifically the precision of the parameters associated with the spatial random effect. The large credible interval about ϕ is likely a result of this disparity, (Table 1). However, the point estimate for ϕ, is consistent with what we might expect in this trial landscape; specifically, high local homogeneity of probability which is independent of land use several thousand meters away.

Methods for multiresolution Gaussian modelling for spatially replicated datasets can potentially be adapted to the logistic model (see e.g., Banerjee and Johnson 2006; Banerjee and Finley 2007). Multiresolution models use the disparity in the concentration of observations to distinguish between *macro*-level and *micro*-level spatial variation. Banerjee and Finley (2007), use multiresolution models with FIA inventory data to describe spatial variation of forest biomass across plots and within plot (i.e., at landscape and local levels).

5.2 Computational considerations

Access to the predictive distribution of each pixel is very useful; however, it comes at a potentially high cost. The algorithms for parameter estimation and prediction are computationally expensive. Depending on the number of observations considered, calculating the inverse of R, which needs to occur in both the Metropolis–Hastings and the subsequent predictions, can be very time consuming. First, for parameter estimation, our original approach was a single block update of $\mathbf{\Omega}$, which required a new candidate ϕ to be drawn and R^{-1} calculation for each block candidate rejection, even if the culprit was among $\boldsymbol{\theta}$, σ^2, and $\mathbf{w}$. Dividing $\mathbf{\Omega}$ into its components and allowing each to be sequentially updated (much like in Gibbs sampling) significantly improved efficiency and, as noted above, provided finer control over acceptance rates.

Second, for any given prediction, R^{-1} must be calculated for each posterior sample, which, depending on the chosen sample size, can make routine mapping of

mid-resolution satellite imagery for even a small area impractical. We partially circumvented this problem by discretizing the posterior samples of ϕ into 1,000 intervals, then setting the sorted vector of unique ϕ as a key in a key-value hash. Then for each key, the associated R^{-1} was calculated and a pointer to this matrix was set at the value component in the hash. This data structure allowed for a fast binary look-up and retrieval of the NR^{-1} required for each prediction, and greatly reduced the time required to map the trial site. Importantly, we found that discretizing ϕ into as few as 100 intervals had negligible effect on the estimates.

Acknowledgments This research was supported by NASA's Earth System Science Graduate Student Fellowship Program. Our manuscript was greatly improved by the many constructive suggestions from Alan R. Ek, three anonymous reviewers, and the Associate Editor.

References

Banerjee S, Carlin BP, Gelfand AE (2004) Hierarchical modelling and analysis for spatial data. Chapman and Hall/CRC Press, Boca Raton

Banerjee S, Finley AO (2007) Bayesian multiresolution modeling of spatially replicated data. J Stat Plann Infer 137:3193–3205

Banerjee S, Johnson GA (2006) Coregionalized single- and multi-resolution spatially-varying growth curve modelling with applications to weed growth. Biometrics 61:617–625

Bechtold WA, Patterson PL (eds) (2005) The enhanced forest inventory and analysis program: national sampling design and estimation procedures. General Technical Report SRS–80. U.S. Department of Agriculture, Forest Service, Southern Research Station, Asheville, NC

Berger J, De Oliveira V, Sanso B (2001) Objective Bayesian analysis of spatially correlated data. J Am Stat Assoc 96:1361–1374

Campbell JB (1996) Introduction to remote sensing, 2nd edn. Guilford Press, New York

Chilés JP, Delfiner P (1999) Geostatistics: modelling spatial uncertainty. John Wiley and Sons, New York

Cressie NAC (1993) Statistics for spatial data, 2nd edn. Wiley, New York

Ferguson TS (1996) A course in large sample theory. Chapman and Hall, London

Franco-Lopez H, Ek AR, Bauer ME (2001) Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. Remote Sens Environ 77:251–1709

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis, 2nd edn. Chapman and Hall, London

Homer C, Huang C, Yang L, Wylie B, Coan M (2004) Development of a 2001 national landcover database for the United States. Photogramm Engi Remote Sens 70(7):829–840

Kauth RJ, Thomas GS (1976) The tasseled cap—a graphic description of the spectral-temporal development of agricultural crops as seen by Landsat. Proceeding of the symposium on machine processing of remotely sensed data. Purdue University, West Lafayett, pp. 41–51

McRoberts RE, Wendt DG, Nelson MD, Hansen MH (2002) Using a land cover classification based on satellite imagery to improve the precision of forest inventory area estimates. Remote Sens Environ 81:36–44

Spiegelhalter DJ, Best N, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). J Roy Statist Soc Ser B 64:583–639

Tomppo E (1991) Satellite imagery-based national forest inventory of Finland. Inter Arch Photogramm Remote Sens 28: 419–424

Author Biographies

Andrew O. Finley is an Assistant Professor with a joint appointment in the Department of Forestry and Department of Geography, Michigan State University. He holds an M.S. in Statistics and Ph.D. in Natural Resources Science and Management from the University of Minnesota. His specific research interests include development of multi-source forest inventory strategies, as well as modeling of longitudinal and spatially correlated data. He is a recipient of the NASA Earth Systems Science Graduate Fellowship.

Sudipto Banerjee is an Associate Professor in the Division of Biostatistics, University of Minnesota. He holds an M.S. and Ph.D. in Statistics from the University of Connecticut, Storrs. His primary research interests include Bayesian modelling in spatial statistics, environmental modelling, public health, and general biostatistics.

Ronald E. McRoberts is a mathematical statistician with the Forest Inventory and Analysis program of the Northern Research Station, USDA Forest Service, in St. Paul, Minnesota. His research interests include nonlinear modelling, model-based estimation of forest attributes using satellite imagery, and spatial uncertainty assessment.